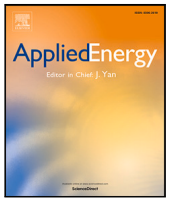
Applied Energy 353 (2024) 122079

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP

Corne van Zyl, Xianming Ye *, Raj Naidoo

Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, 0002, South Africa

ARTICLE INFO

Keywords:
Explainable Artificial Intelligence (XAI)
Energy forecasting
Gradient-weighted Class Activation Mapping (Grad-CAM)
Shapley Additive Explanations (SHAP)
Convolutional Neural Network (CNN)
Feature selection

ABSTRACT

This study investigates the efficacy of Explainable Artificial Intelligence (XAI) methods, specifically Gradient-weighted Class Activation Mapping (Grad-CAM) and Shapley Additive Explanations (SHAP), in the feature selection process for national demand forecasting. Utilising a multi-headed Convolutional Neural Network (CNN), both XAI methods exhibit capabilities in enhancing forecasting accuracy and model efficiency by identifying and eliminating irrelevant features. Comparative analysis revealed Grad-CAM's exceptional computational efficiency in high-dimensional applications and SHAP's superior ability in revealing features that degrade forecast accuracy. However, limitations are found in both methods, with Grad-CAM including features that decrease model stability, and SHAP inaccurately ranking significant features. Future research should focus on refining these XAI methods to overcome these limitations and further probe into other XAI methods' applicability within the time-series forecasting domain. This study underscores the potential of XAI in improving load forecasting, which can contribute significantly to the development of more interpretative, accurate and efficient forecasting models.

1. Introduction

Accurate short-term load forecasts are essential to the optimal management of energy production and consumption. Load forecasting can be challenging due to the ever-growing demand of an expanding economy, changes in weather conditions, shifts in consumer behaviour, and rapid advancements in technology like solar panels and electric vehicles. These factors can alter the energy landscape in unpredictable ways [1]. When these forecasts are applied to utility companies, small improvements in accuracy can have significant financial benefits [2], thus the pursuit of higher load forecasting accuracy is worthwhile. In recent years, artificial intelligence (AI) and machine learning (ML) techniques have been applied to energy forecasting with promising results [3,4], but a major challenge remains in the lack of transparency and interpretability during the energy forecast process. It is possible to examine the internal computations of simpler models, such as linear regression or decision trees, to uncover the reasoning behind the model predictions. As moving towards more complex model structures like deep neural networks, which can have millions of parameters, interpreting the computations becomes an endeavour far beyond human understanding. The inability to understand ML models ultimately makes it difficult to trust the output of the models [5].

The eXplainable Artificial Intelligence (XAI) focuses on making ML models more comprehensible and transparent. The XAI methods provide interpretability to machine learning models by mapping abstract concepts learned by the model to a domain that can be understood by humans [6]. A popular approach in XAI is to identify feature attributions, which are scores assigned to each input of the model that quantifies their significance to the prediction made by the model. Consequently, by discerning the features that have a significant impact on the model, the output produced by the XAI can be interpreted as an explanation for the given prediction. Although XAI has been used primarily to understand the decision-making processes employed by the ML models, it has also been demonstrated to be an effective feature selection tool [7]. The rationale behind this approach is that features with low attribution scores are minimally used by the model and can thus be removed without significant adverse effects to the model performance.

As it is difficult to know in advance which features should be included in the model, users tend to include all the available features, believing that the model training process will decide which features are important [8]. It is common for datasets with a large number of features to contain irrelevant ones [9], and including nonessential features that may negatively impact the performance of the model. Thus, a process is required to effectively select and retain only the most relevant features. This process, known as feature selection, involves identifying and retaining a subset of significant features that preserves high predictive performance, while removing the less relevant

* Corresponding author.
E-mail address: xianming.ye@up.ac.za (X. Ye).

https://doi.org/10.1016/j.apenergy.2023.122079
Received 22 June 2023; Received in revised form 5 September 2023; Accepted 2 October 2023
Available online 17 October 2023

ones [10]. Feature selection exhibits the ability to enhance model accuracy, improve computational efficiency, and tend to make models more interpretable [11,12]. When dealing with datasets that contain a multitude of features, implementing feature selection is essential to mitigate computational load, especially during the iterative process of model design and improvement. This makes tasks like hyperparameter tuning of ML models more feasible. Furthermore, the enhanced interpretability is vital in high-risk applications where the rationale behind predictions is as important as the accuracy of the predictions themselves. This is the realm where XAI proves invaluable in facilitating an understanding of the decision-making processes within the models.

Extensive advancements have been made in the domain of XAI, however its application in multivariate time series forecasting models is largely unexplored, especially as a feature selection tool. Time-series data, with its inherent chronological ordering and numerous interacting variables, is often high-dimensional and non-intuitive. In such cases, incorporating XAI could not only enhance the interpretability of the model but also aid in feature selection by identifying which variables in these high-dimensional datasets have the most influential contribution. Existing literature has shed light on the capacity of XAI to reveal the inner working mechanisms of ML models. Evaluating the quality of the explanations provided by different XAI methods remains a challenging task [13]. Our study aims to assess the potential of XAI as a tool for feature selection, in an effort to improve the performance of a deep learning load forecasting model. We undertake a comparative analysis of Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) by examining their ability to systematically rank predictive features based on their contributions to forecast accuracy. We start by adapting Grad-CAM and SHAP for use with multi-headed CNNs, thereby broadening their compatibility with a wider range of deep learning architectures. Subsequently, we employ Grad-CAM and SHAP to assign feature attributions, first to the inputs of a single prediction, offering in-depth insight into the XAI methodologies. Finally, we evaluate the feature rankings through an ablation study to observe the performance of the model when low-attribution features are gradually removed. Although the application of XAI techniques for feature selection is still in its infancy, our analysis of Grad-CAM and SHAP provides valuable insights into their respective strengths and limitations, thereby informing both their future refinement and broader applicability.

2. Related studies

2.1. Traditional feature selection

Feature selection methods generally fall into three categories: filters, wrappers, and embedded methods [11]. Filter methods evaluate feature importance based on the characteristics of the data, such as statistical properties between features and the predicted variable [14]. They often utilise metrics such as Euclidean distance, Pearson correlation, and mutual information, to name a few [14]. As they do not depend on the use of any learning algorithm, filter methods are generally more computationally efficient and scalable to high-dimensional datasets than wrapper methods. However, the filters do not account for interactions with specific models, potentially leading to feature sets that are not optimised for a particular model [15]. Studies have found that filter methods generally perform poorly compared to wrapper and embedded methods [16].

Wrapper methods wrap the feature selection around a learning algorithm and use the accuracy of the algorithm to evaluate the predictive power of the feature set [17]. Feature selection is thus turned into an optimisation problem where a search algorithm is employed to find the subset of features that maximises the performance of the learning algorithm. As a result, wrapper methods often yield better accuracy than filter methods. However, the wrapper-based approach requires repeatedly training the learning algorithm to assess the effectiveness of various feature combinations. Consequently, they can be computationally demanding [10]. The computationally demanding nature of wrapper methods poses a challenge to their practicality and effectiveness when implemented with complex models and high-dimensional datasets [18]. One popular strategy to reduce the computational load is to use a less complex, and hence more computationally efficient model for feature selection that allows for quicker training times. For instance, researchers in [19] used an extreme learning machine (ELM) combined with particle swarm optimisation (PSO) and genetic algorithm (GA) to select the optimal subset of features for a day-ahead electricity price prediction model. Subsequently, the proposed long short-term memory (LSTM) deep neural network, trained using the optimal feature subset, exhibited superior performance. Note that when employing simpler models for feature selection, the effectiveness of chosen features may not be consistent across different machine learning models [20]. For a more comprehensive exploration of wrapper methods and their applications, readers can consult [21–23].

Embedded methods exploit the properties of the learning algorithm, where feature importance is obtained directly from the patterns learned by the model [11]. They are implemented by algorithms that possess their own built-in feature selection methods [10]. For instance, Decision Trees and derivative algorithms like Random Forest or Gradient Boosting allow for measuring feature importance directly from the model [24,25]. The Least Absolute Shrinkage and Selection Operator (LASSO) forces some of the coefficient estimates to be exactly zero when the regularisation parameter is sufficiently large [26]. Features with zero coefficients are effectively removed from the model, thereby performing feature selection.

Hybrid methods integrate various feature selection techniques to harness the strengths of each approach. Combining a fast method with a slower, more thorough one can strike a balance between computational efficiency and feature selection quality. For example, in a study focused on multivariate financial time series forecasting, the researchers first employed RReliefF [27] for preliminary feature reduction. Subsequently, they identified the optimal feature set using a wrapper approach, which combined a multi-objective binary grey wolf optimiser with Cuckoo search and ELM [28]. Another popular method, recursive feature elimination with support vector machines (RFE-SVM), combines both embedded and wrapper approaches [29]. The RFE-SVM involves training an SVM on all features, ranking them by the magnitude of their weights, and then recursively eliminating the least important features. Although RFE-SVM is effective in identifying valuable features, it can be computationally intensive because the SVM model must be retrained each time a feature is eliminated.

2.2. XAI

The XAI methods for feature selection do not strictly fall into the traditional categories of the filter, wrapper, or embedded methods. Instead, they constitute a different kind of approach that we could term ‘‘interpretability-based methods’’. The XAI methods share similarities with the traditional methods; like embedded methods, feature importance is obtained from the trained model and does not require iterative training of the model. However, the goal of XAI methods extends beyond merely improving prediction accuracy. They also aim to provide insights into the model’s decision-making process. The ability to understand the contribution of each feature to a prediction can guide the feature selection process, highlighting those features that are truly significant versus the features that are insignificant. Early model interpretation methods are detailed in [6]. Since then, the diversity of XAI methods has expanded significantly, finding application in time series classification [30], as well as being utilised extensively within the power and energy systems domain [5].

XAI methods can broadly be divided into two categories: model-specific methods and model-agnostic methods. Model-specific methods are interpretability techniques tailored to specific types of models. They

leverage the intrinsic structure of a particular type of model to provide insights about the learned relationships. Model-agnostic methods are designed to interpret any machine learning model, regardless of its internal working mechanism or structure. Model-agnostic methods aim to make the predictions of any machine learning model understandable, irrespective of the complexity.

The output of XAI methods, often termed as ‘‘explanations’, varies in scope. The local explanations generated by the XAI methods focus on understanding why the model made a particular prediction for a single example or instance in the dataset. For instance, in a load forecasting model, a local explanation focuses on why the model predicted high energy consumption at a specific time instance. In contrast, global explanations attempt to provide a broad understanding of the model's general prediction process across all predictions. For instance, it aims to highlight which features, like time of day or weather conditions are generally most important for the model's energy predictions.

Grad-CAM, which stands for Gradient-weighted Class Activation Mapping, is a model-specific XAI technique designed to work with deep learning models that contain Convolutional Neural Networks (CNNs). Grad-CAM exploits a property of CNNs where the spatial relationship in the data is maintained after it passes through the convolutional layers [31]. Multi-layer perceptrons (MLP), also known as fully connected layers, do not preserve the spatial relationships within the data, making it challenging to map important regions back to their original locations in the input. The convolution operation applies filters or kernels that slide over the input to generate feature maps. Each filter is trained to recognise and respond to a specific pattern in the input and highlights the areas in the input where that pattern is present. The sensitivity of the model's output to minor changes in the feature map determines the importance of that feature map. By collecting the weighted contribution of the feature maps, Grad-CAM is able to highlight features that are important to the model's prediction.

Originally designed to explain the outputs of image classification models, Grad-CAM has been applied in time series forecasting, notably to explain the predictions of a residential energy consumption model that employs a CNN-LSTM architecture [32]. By selecting a 2D CNN the authors in [32] treated the batched multivariate time series data as an image. Grad-CAM is not exclusively used with 2D CNNs. It has been effectively applied to 1D CNNs as well, shedding light on key contributors to Heating, Ventilation, and Air Conditioning (HVAC) faults [33]. Grad-CAM has been applied to time series classification models, used to explain the predictions of a power quality disturbance classifier by highlighting the regions in the signal that led to the classification [34]. The researchers behind MTEX-CNN [35] introduced a combined approach using both 2D and 1D CNNs to generate spatio-temporal explanations with the help of Grad-CAM. The 2D CNN's role is to assign importance to individual features, while the 1D CNN is used to determine the overall temporal significance over time. This dual-layered structure enables a comprehensive understanding of both spatial and temporal aspects, providing a more complete picture of the features' importance during modelling processes. It is noted that if only a 1D CNN is used for the multivariate time series, it would combine the influence of all the different features at each time step into a single value. This means that a 1D CNN would not be able to distinguish the importance of specific features at each time step. Instead, it would highlight which time steps are important overall, taking into account all features. Hence, the study [35] chose to use 2D CNNs. Diverging from the conventional use of Grad-CAM as a local explainer, the study in [35] extended its application to generate predictions across the entire test dataset. By collecting positive feature attributions from each forecast, they effectively use Grad-CAM as a global explainer. The reliability of the global feature attributions is confirmed by observing a minimal impact on prediction accuracy when low-attribution features are excluded from model training. Grad-CAM was effectively used to identify which channels of the electroencephalography (EEG) are most informative for classifying different intentions or mental states [36]. By applying Grad-CAM as a feature selection tool, the model maintained a decoding performance of 92.31% while reducing the number of channels by nearly half, thereby improving the decoding rate of the system.

In [37], an eXplainable convolutional neural network for Multivariate Time Series (MTS) classification extends the work of [35] by proposing the use of 2D and 1D CNNs within a parallel architecture with the purpose of achieving high-resolution explanations. The study [37] suggest that the necessity for upscaling, an operation performed when the size of the feature maps created by the CNN is smaller than the original input, can lead to a reduction in the resolution of the ensuing explanations. The term ‘‘resolution’’ refers to the precision in identifying key features in the time series that are most influential in the prediction.

Local Interpretable Model-agnostic Explanations (LIME) [38] is a model-agnostic XAI method, which has been popular in numerous classification tasks. However, its applicability to time series models has been called into question. Studies have shown that both the classic implementation and its time series implementation, LimeforTime, struggle to identify key features or time steps that contribute significantly to model predictions [34,39,40].

Shapley values, which originate from cooperative game theory, are adopted to quantify each player's individual contribution to a game. The Shapley values facilitate the fair distribution of the total gain generated by a game among its players based on their contributions. In machine learning, the SHapley Additive exPlanations (SHAP) approach [41] views the model's prediction for a given instance as the game and the features used in the model as the players. Previous studies have applied SHAP to explain time series models [42,43]. The application of SHAP in these studies does not extend to deep learning models. The study by [44] includes comparisons to deep learning models, however they, along with other studies [42,43,45,46], perform what we refer to as mixed multivariate time series forecasting. Mixed multivariate time series forecasting uses a combination of time series and singular values to represent exogenous variables to predict future values of the target variable. In contrast, pure multivariate time series forecasting involves predicting a target variable solely based on the values of multiple time series variables. This allows the model to learn from the interdependencies between different time series variables, capturing complex temporal patterns that might be missed when expressed as only singular values. However, the application of SHAP in deep learning models for pure multivariate time series forecasting has not been extensively explored. The study [39] uses a deep learning model, ResNet, to classify maritime traffic by analysing the multivariate time series data of vessels. The study [39] conducts a comparative analysis of several model explanation techniques including DeepSHAP and GradientSHAP, both variants of SHAP, along with Path Integrated Gradients (PIG), as well as LimeforTime. In [39], it is found that SHAP variants produced the best explanations, whereas LimeforTime performed the worst in identifying the key features. Although the exact methodology used to generate SHAP values is not explicit, speculation based on their use of ResNet, a model that utilises 2D CNNs, suggests that the authors may have used the same approach as [32] and treats the multivariate time series as an image. Furthermore, the SHAP documentation [47] includes examples of its application to images using the ResNet architecture, which reinforces this hypothesis.

Despite the merits of SHAP, there are some notable caveats. Given that the SHAP's feature attributions are relative to a baseline, the choice of this baseline significantly influences the resulting explanations [48]. The intricacies of choosing a suitable baseline are explored in detail in the works of [49,50]. Moreover, some researchers argue that SHAP is not inherently designed for time series models as its sampling methodology may generate instances that violate the temporal ordering of the data [51]. As a result, the model could be perturbed by unseen data, leading to excessively large responses and ultimately produce artificially inflated attributions. The study in [52] argues that while

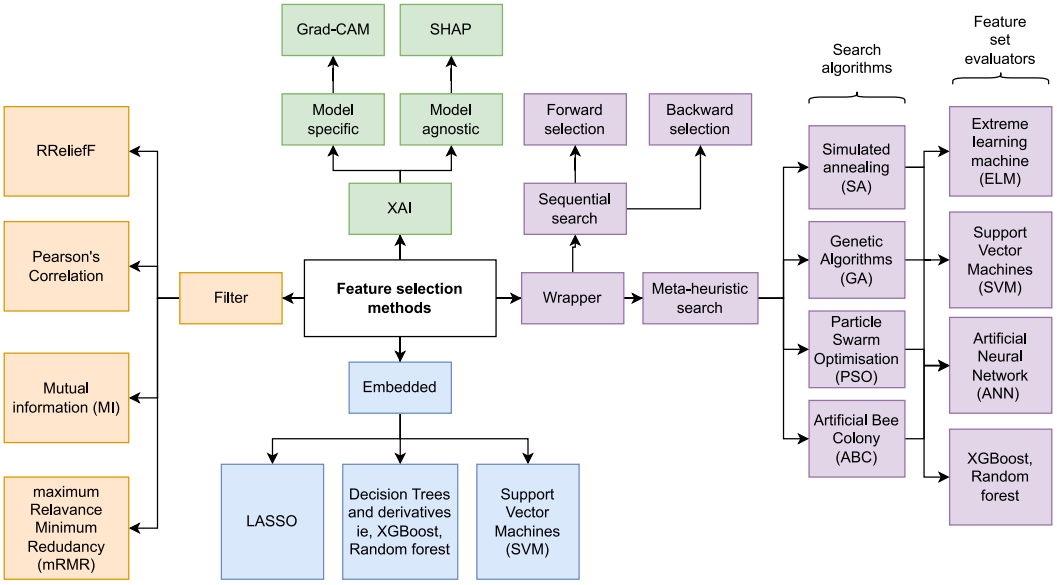


Fig. 1. Overview of feature selection methods.

the Shapley value is theoretically sound, it may not always align with the goals of feature selection in XAI. This is due to its axioms, such as model averaging, which may not accurately reflect the performance of features in optimal submodels or account for interactions with dominant features.

Despite the criticisms of SHAP, it has been demonstrated to select features more effectively than ANOVA, Mutual Information, and Recursive Feature Elimination [53]. Additionally, in a study comparing filter, wrapper, and embedded methods on environmental data, SHAP consistently excelled in terms of stability and overall efficacy [16]. It is worth noting that the study in [16] did not compare any swarm intelligence based wrapper methods which leaves doubt as to how SHAP performs to the most accurate feature selection methods.

2.3. Feature selection in load forecasting

Various methods, including filter techniques, have been employed to enhance the prediction accuracy of load forecasting models. In [54], it implemented a two-stage mutual information feature selection technique (MIT-MIT). In [55], the minimal redundancy maximum relevance (mRMR) method was used. Both studies demonstrated improved forecasting accuracy compared to models without feature selection. Yet, without comparative analysis against other techniques, assessing the optimality of their chosen feature sets becomes challenging. Research on multivariate time series energy forecasting for buildings found that wrapper methods outperformed filter methods, though they also highlighted the high computational costs associated with the former [56]. In [15], IT used a wrapper-based approach with an artificial neural network (ANN) for short-term demand prediction. Alongside this, they incorporated a binary genetic algorithm (GA) to search for the most effective feature set. While the GA-ANN approach was effective when applied to an ANN with 24 features, its suitability for datasets with a larger number of features remains uncertain. This is because the complexity of the ANN increases with the number of features, which in turn makes the training process and the evaluation of the features more time-consuming. In [57], a hybrid feature selection method was used for industrial load forecasting. The approach reduced the 44 initial features using Gradient Boosting Decision Trees (GBDT) and Pearson correlation by removing those with low contribution or high correlation.

Previous studies have also utilised XAI for feature selection in load forecasting. A study by [58] utilised SHAP to analyse the influence of features on predictions, grouping them according to their Shapley values. The performance of the selected features was then validated through an ablation study, in which an LSTM model was retrained using only those selected features. The study [58] found that models with high Shapley value features ensured high forecasting accuracy, while those with low Shapley values underperformed. However, the study [58] is limited to models that perform single-step forecasts and use mixed multivariate time series data, shaped into a 1D vector, as input.

In summary, Fig. 1 provides a detailed visual summary of the diverse feature selection methods discussed.

2.4. Energy forecasting models

CNNs are known for their ability to extract local features and patterns from input data while maintaining spatial invariance, which has made them a popular choice in image classification models [59]. However, the CNN application to time series has not been as widely acknowledged compared to Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. Studies employed CNNs for energy forecasting have reported considerable success. For instance, [60] demonstrated that 1D CNNs outperformed LSTM and GRU networks in predicting hourly electricity load for the city of İstanbul. Similarly, a study comparing the performance of different CNNs for national peak demand forecasting found that 1D CNNs provide superior forecasting accuracy compared to other deep learning models, including LSTM [61]. Interestingly, the study [61] also reported that multi-headed CNNs ranked second, closely following 1D CNNs, while 2D CNN-LSTM and LSTM often failed to model the demand accurately. These promising results underscore the potential of CNNs for time series forecasting, especially in the energy sector.

2.5. Contribution

In summary, filter methods, known for their scalability, operate independently of machine learning algorithms. While they have been shown to improve forecasting accuracy, they are generally considered to perform poorly compared to wrapper and embedded methods. While wrapper methods often produce more acceptable results, they are computationally demanding and might be infeasible for multivariate time series. Furthermore, wrapper methods only provide the user with an optimal feature set and do not offer insight into how the features affect

the model. Embedded methods, on the other hand, show promise as an effective feature selection tool for high-dimensional datasets, which typically reveal only the magnitude of a feature's significance, not the specific nature of its influence on the model.

XAI offers computational benefits for feature selection by eliminating the need for iterative model training. Additionally, given that features are selected based on patterns learned by the model, feature interactions, temporal dependencies, and how features interact with the model are accounted for. Lastly, XAI methods furnish insights into the model's decision-making process, shedding light on the specific influence exerted by each feature. However, challenges remain in the application of XAI to various types of models and data. In the literature, there appears to be a significant lack of diversity in the deep learning model architectures explained using SHAP. The application of SHAP has been largely limited to either ANNs or LSTM networks, with the notable exception of the study cited in [39]. Furthermore, there is lack of research on the application of Grad-CAM to multivariate time series forecasting models. Those found in literature only apply to classification tasks [35–37].

Our key contributions are as follows:

1. Our study is the first to directly compare the utility of two XAI techniques, namely Grad-CAM and SHAP, in the context of feature selection in load forecasting with multivariate time series data.
2. We adapt Grad-CAM and SHAP for use with multi-headed CNNs, making them compatible with a wider range of deep learning architectures. Specifically, we demonstrate that by adding a layer into the model pipeline that performs the necessary array manipulations, SHAP can be adapted to suit diverse model architectures. To the best of our knowledge, our study is also the first to use Grad-CAM as a feature selection tool for multivariate time series load forecasting.
3. We perform a comparison of Grad-CAM and SHAP in their ability to rank predictive features based on their contributions to forecast accuracy. Through an ablation study, we assess how the removal of low-attribution features affects model performance, thereby validating the feature rankings provided by the XAI methods.
4. Our research offers a nuanced understanding of both the strengths and limitations of Grad-CAM and SHAP. This insight is invaluable for assessing the suitability of these methods for specific applications. Furthermore, it serves as a roadmap for enhancing these techniques and for exploring alternative approaches in the field of XAI.

The multi-headed CNN was chosen for this study for several strategic reasons. First, it allows for compatibility with Grad-CAM. Given that Grad-CAM can only be applied to models that contain a CNN, it excludes models such as LSTM. While 1D CNNs are effective for the task at hand, they have limitations when used with Grad-CAM, particularly in attributing importance to individual features in multivariate time series data. Secondly, the architecture is designed to inherently account for the varying importance of different features. Since not all feature maps generated by the CNNs within each head of the model contribute equally to the prediction, Grad-CAM can effectively capture these differences in significance, thereby enabling the assignment of attribution scores to individual features. Thirdly, although 2D CNN-based models, as cited in [32,39], could have been suitable for our study, the multi-headed CNN serves as a versatile platform that showcases the flexibility of our methodology, enabling the application of both Grad-CAM and SHAP. This is particularly noteworthy given the limited diversity in deep learning architectures that can be explained using SHAP, as observed in existing literature. Given that SHAP is widely regarded as one of the most prevalent methods in the field of XAI, we believe our adaptation has the potential for significant impact within this domain. Finally, it is worth noting that the primary aim of our study is not to devise the most accurate forecasting model per se. Nonetheless, the multi-headed CNN has exhibited competitive performance as evidenced by previous research [61], making it a suitable choice for our investigative purposes.

3. Methodology

The methodology begins with the pre-processing and transforming of the original tabular time series data into batch series data. This process, described in detail in Section 3.1, allows for the transformed data to be utilised for subsequent model training and testing. Following data preparation, we outline the design and evaluation criteria of the multi-headed CNN in Sections 3.2 and 3.4. Local and global explanations for individual and multiple test samples are produced using Grad-CAM and SHAP. The feature importance scores from the multiple samples are then combined to generate an overall ranking of the significance of each feature. The intricacies of generating feature attributions are laid out in Section 3.3. Finally, the rankings of feature importance are assessed through an ablation study and then compared to validate their effectiveness.

3.1. Data preparation

Data enrichment and augmentation

Date and time information is extracted and divided into separate columns, which are represented as integers for the hour, day of the week, day of the month, month, and year. Each holiday throughout the year is denoted as an integer, and a binary representation is included. Sinusoidal transformations are applied to periodic features x_c, such as the hours of the day, days of the week, days of the month, months of the year, and holidays of the year.

$$x_{c,j}(i) = \cos\left(\frac{2\pi x_j(i)}{P_j}\right), \tag{1}$$

where P_j is the period of the series, i is the index an individual data point within the series of feature j. First-order differences of all time series variables are incorporated to explore the effects of the rate of change of these variables. To improve forecast performance, lead series for calendar events and weather data are shifted forward, providing insights into upcoming time intervals and expected weather conditions that may affect energy consumption. Lead series are found by,

$$x_{lead,j}(i) = x_j(i+T), \tag{2}$$

where T is the length of the load forecast horizon.

Data preprocessing

The dataset undergoes normalisation, where each feature's time series is transformed to have a mean of zero and a standard deviation of one. This transformation can be represented by

$$x_{std,j}(i) = \frac{x_j(i) - \mu_j}{\sigma_j}, \tag{3}$$

where μ_j, σ_j, and $x_j(i)$ correspond to the mean, standard deviation, and the value at the ith time step of the jth feature, respectively.

Sliding window algorithm

After scaling the data, we apply the sliding window algorithm. This technique, commonly used in time series analysis, involves moving a fixed-length window along the time series, advancing one data point at a time. This procedure transforms the tabular dataset into a batched series. The resultant three-dimensional array (Q, N, J) consists of Q batched samples of (N, J), where J is the number of features in the dataset and N is the length of each series. The final step involves transforming the batched time series into a 2 dimensional, tabular array $(Q, (N \times J))$ to prepare it for use in the SHAP algorithm. This is done by transforming each (N, J) sample into a 1D array $((N \times J), 1)$, by concatenating the end of one series to the start of another.

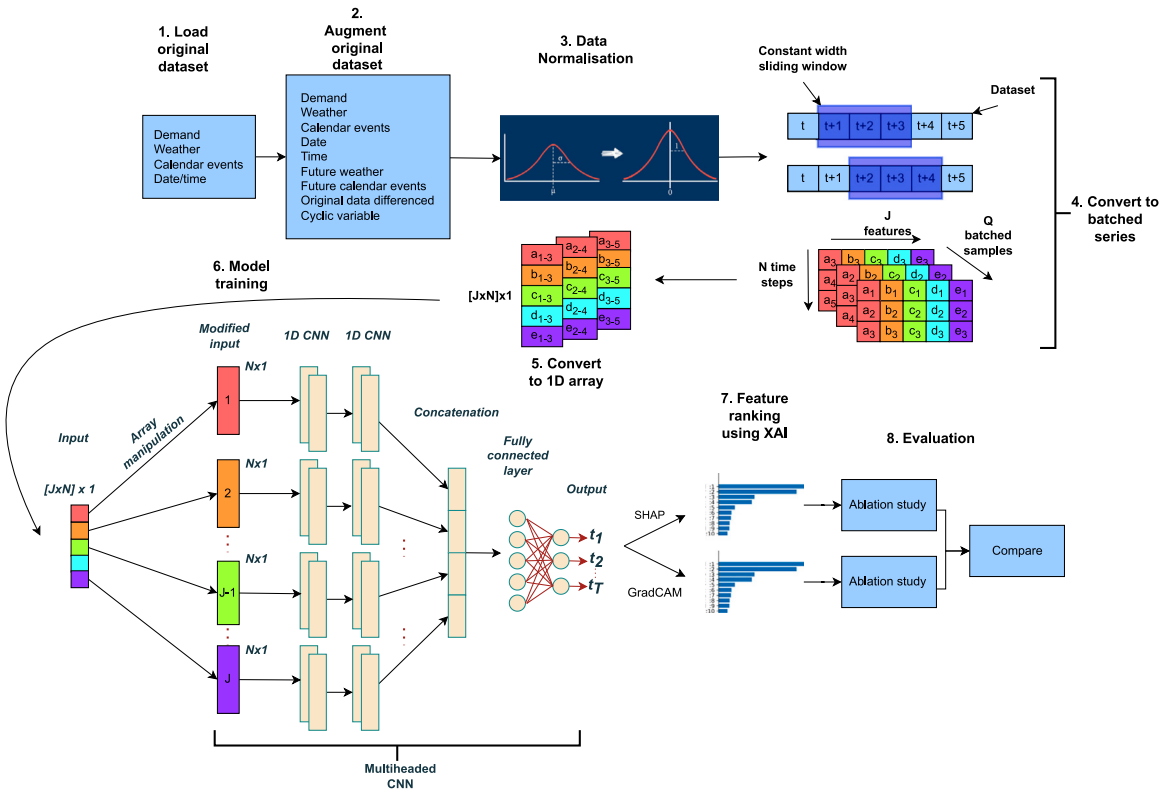


Fig. 2. Methodology overview.

3.2. Model design and setup

The existing SHAP tool for tabular datasets, operates by processing a single row of data from the input table at a time and thus requires the model to accept only one-dimensional inputs. This creates a problem as state-of-art deep learning architectures, such as those in [35,37,62], do not conform to this requirement. This restricts the architectures that can be explained using SHAP which might contribute to the lack of diversity in deep learning architectures observed in the literature. This constraint can be overcome by first transforming the batched time series samples (Q, N, J) into a tabular format $(Q, (N \times J))$, as done in previous studies [42,43,45,46], and designing the deep learning model to it accepts a 1D input vector. An arbitrary architecture can be explained using SHAP, by adding a layer to the model's architecture, after the input layer, that manipulates the shape of the input array to the required shape that is compatible with desired architecture.

The multi-headed CNN as illustrated in Fig. 2, is designed with this solution in mind. It accepts a one-dimensional array $((N \times J), 1)$, transforms it into a two-dimensional array (N, J), and assigns each feature's time series $(N, 1)$ to a corresponding head in the architecture. The number of heads in the model thus corresponds to the number of time-series variables in the input data. The model consists of multiple parallel 1D CNNs, with each head containing two sequential 1D CNNs. Notably, no pooling operation is performed after any of the CNNs, and padding is used to maintain the dimensions of the feature maps. The output of the last CNN in each head is concatenated and fed into a fully connected layer. The final layer is another fully connected layer that generates the predictions, with the number of neurons in this last layer being dictated by the number of steps in the prediction.

Given the large input dimensions and model parameters, hyperparameter optimisation is computationally intensive. Instead, we use a trial-and-error approach, incrementally increasing network sizes until further accuracy improvements become marginal. To prevent overfitting during model training, we implement early stopping, stopping training when the performance on the validation set plateaus. This strategy obviates the need to compare training and testing metrics or determine the optimal number of training epochs. Given that multi-headed CNN's internal parameters initialise with random values at the start of the training process, we train 10 models for each experiment to capture the variance in model performance.

3.3. XAI methods

In this section, we apply and compare the SHAP and Grad-CAM explainability techniques in the context of load forecasting. We detail how each method assigns attributions scores to features and discuss their application to the multi-headed CNN.

3.3.1. SHAP

The fundamental idea underpinning SHAP is the assignment of a score to each feature in a given data point that quantifies the degree to which that feature contributed to the model's prediction. These scores are based on the Shapley values from cooperative game theory, which is an approach to attributing the value of a collective effort to each individual participant. The exact Shapley values ϕ_j for feature j are computed as follows:

$$\phi_j = \sum_{S \subseteq X \setminus j} \frac{|S|!(|X| - |S| - 1)!}{|X|!} \left[f(x_j) - f_S(x_j) \right], \tag{4}$$

where X is the set of input features, $S \subseteq X \setminus j$ is a subset of features that does not include feature j, $f(x_j)$ is the model's prediction for the input sample x_j, $f_S(x_j)$ is the model's prediction for the input sample x_j with the features in S set to their expected values.

Computing the exact Shapley values requires 2^J calculations, which is impractical in practice. Therefore, approximation techniques are used to estimate Shapley values. The SHapley Additive exPlanations (SHAP) [41] method provides an efficient way of estimating the Shapley values for a specific model and dataset and while not exact, has shown its efficacy in real-world applications and machine learning interpretability research. KernelSHAP, GradientSHAP, and TreeSHAP

are some of algorithms that estimate Shapley values.[1] In this study, DeepSHAP is utilised, which is specifically designed for deep learning models.

The SHAP algorithm generates an attribution score for each feature in the model's input, ϕ_j, which quantifies its contribution to each output of the model. The fundamental notion is that the sum of all feature contributions approximates the difference between the baseline ϕ_0 and the prediction $f(x)$ being explained.

$$f(x) \approx \sum_{j}^{N} \phi_j(x) + \phi_0, \tag{5}$$

where each $\phi_j(x)$ represents the contribution of feature j towards driving the model's output away from its baseline value ϕ_0 for the given input x. The baseline output of the model over all inputs, ϕ_0, can be calculated by taking the average of the model's output over the entire input distribution

$$\phi_0 = \mathbb{E}_x[f(x)]. \tag{6}$$

The baseline, also called the expected or average SHAP value, represents the average model prediction when all features are considered absent [41]. In our study, the we randomly select a thousand samples from the whole dataset to represent the baseline. Additionally, we draw a hundred random samples from the test set from which global feature attributions are obtained.

The SHAP algorithm produces feature attributions for all Q samples that relate the inputs of the model ($J \times N$) to all the T outputs of the model. Thus, the output of the algorithm produces a 3D array $(Q, J \times N, T)$. The SHAP values are transformed to (Q, J, N, T) to identify which set of attributions belong to which set of inputs. Finally, the global feature importance $\boldsymbol{\Phi}_j$ for the jth feature is the total impact of the feature taken over the length of the series N, across all Q samples and for all T the model outputs, given by

$$\boldsymbol{\Phi}_j = \frac{1}{Q} \sum_{n}^{N} \sum_{q}^{Q} \sum_{t}^{T} \left| \phi_{j,q,n,t} \right|. \tag{7}$$

3.3.2. Grad-CAM

The Grad-CAM for 1D CNNs is obtained using the same process as for 2D CNNs, but with one dimension less. A forward pass is carried out in the model, which essentially processes the input data through the network's successive layers until an output is obtained, represented as y. During this process, the model also generates a series of feature maps A^k. The total number of chosen hyperparameter that determines the number of filters or kernels in the CNN is noted as K. These feature maps are outputs of the CNNs that capture and represent different learned patterns from the input data. For Grad-CAM, we are only interested in the feature maps produced by the last CNN layer in the model that is closest to the output. The last CNN layer is chosen as it is able to capture the most complex and abstract patterns in the data [31].

At this juncture, our method departs from the strategies employed in classification-based models, in which gradients are computed with regard to a particular class, or to put it differently, a singular output. Given that our forecast model predicts multiple future time points, our objective is to comprehend the influence of each feature across all these predictions, rather than focusing on a single future time point. Thus, we take the gradient of all the outputs with respect to each feature map. The feature attributions Ω_j for the jth feature are represented by the weighted sum of the feature maps, calculated using

$$\Omega_j = ReLU\left(\sum_{k=1}^{K} w_j^k A_j^k \right), \tag{8}$$

where the weights w_j^k of the jth feature represent the importance of each feature map A_j^k. The rectified linear unit (ReLU) function ensures that only positive contributions are considered. As A_j^k has already been found in the forward propagation of the model, the weight w_j^k of the kth filter is found by computing the gradient of all the output values with respect to each feature map, given by

$$w_j^k = \frac{1}{N} \sum_{i=1}^{N} \max\left(0, \frac{\partial y}{\partial A_{i,j}^k}\right). \tag{9}$$

This method sums up all the positive gradients at index i of the feature map A_j^k and averages them based on the number of elements in the feature map N. The strategy of considering only positive gradients is reported to offer superior resolution and emphasise the most consequential input components to the prediction [31]. Given that we use padding in the CNNs, feature map A_j^k and jth input series x_j have the same dimensions. Deep learning frameworks like TensorFlow and PyTorch have built-in functionalities for automatic differentiation, which simplifies the computation of these gradients. To obtain the computational benefits claimed in this study, the feature maps of each head must be obtained together in the same forward pass of the model. Attribution scores for each feature are obtained by iteratively calculating Eqs. (8)–(9).

In [35], global feature attributions, which are also referred to as average feature importance, are derived by making predictions for the entire test set of 365 non-overlapping samples, and then combining the feature attributions of the individual forecasts.

3.4. Performance evaluation

The coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) are used to evaluate the accuracy of the forecasting model,

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2}, \tag{10}$$

$$MAPE = \frac{100\%}{T} \sum_{t=1}^{T} \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \tag{11}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2}, \tag{12}$$

where T is the number of predictions, y_t is the actual value of the tth observation, $\hat{y}_t$ is the predicted value of the tth observation, and $\bar{y}$ is the mean of the observed values.

Given that feature attributions are obtained using the patterns learned by the model, their credibility is inherently tied to the model's proficiency in fitting the data. To ensure robust feature rankings via XAI methods, we train the multi-headed CNN ten times on all features, selecting the top-performing model to generate the rankings. By applying both Grad-CAM and SHAP on the same model, a fair comparison of their performance is ensured. The evaluation of these feature rankings is conducted through a feature ablation study. This technique involves training a model using all features initially, then removing features based on its importance as ranked by the XAI method. It is anticipated that as more features are removed, particularly those ranked as important by the XAI method, there will be a corresponding decrease in the model's predictive accuracy. A greater decrease in accuracy is indicative of greater feature significance.

4. Case study

Panama, strategically positioned in Central America, serves as a land bridge connecting North and South America. The country's climate is predominantly tropical, characterised by high temperatures and humidity. In this study, we utilise a dataset that captures the hourly national demand of Panama. This dataset also includes weather conditions from three major cities: Tocumen, which is near the capital, Panama City, home to the Panama Canal and a significant hub for

[1] https://shap-lrjball.readthedocs.io/en/latest/index.html

Table 1
Description of dataset variables.

Description	Feature name	Unit/range
Demand	Demand	MW
Weather features		
Temperature	Temp - Toc, San, Dav	°C
Specific humidity	Hum - Toc, San, Dav	kg/kg
Precipitation	Rain - Toc, San, Dav	liters/m^2
Wind speed	Wind - Toc, San, Dav	m/s
Calendar features		
Holiday ID	Hol ID	[1–22]
Holiday indicator	is hol	[0, 1] Holiday = 1
School holiday	is scho	[0, 1] School = 1
Weekend indicator	isweekend	[0, 1] Weekend = 1
Date and time features		
Hour of day	hour	[0–23]
Day of the week	dayofweek	[0–6] Monday = 0
Day of month	day	[1–31]
Month of the year	month	[1–12]
Year	year	[2015–2019]
Augmented features		
Lead features	Future - [Weather, Calendar]	
Differenced features	diff - [Demand, Weather , Calendar, Date and Time]	
Sinusoidal features	cyclic - [Hol ID, hour, day, dayofweek, month]	[1 to −1]

Toc = Tocumen, San = Santiago city, Dav = David city

Table 2
Multi-headed CNN architecture setup and parameters.

Layer	Parameters
Input_j	shape = [1 × 168]
1D CNN_(0,j)	filters = 16, kernel size = 2, padding = same, activation = ReLU
1D CNN_(1,j)	filters = 16, kernel size = 2, padding = same, activation = ReLU
Concatenation	axis = −1 (default)
Dense_0	neurons = 200, activation = ReLU
Dense_1 (Output)	neurons = 24

commerce; David City, the third-largest urban centre; and Santiago, a centrally located key regional centre influencing both urban and rural demands. Additionally, the dataset incorporates public and school holidays, which can impact energy demand patterns [63]. The dataset spans from 3 January 2015 to 27 June 2020. However, we exclude data from 2020 because the COVID-19 pandemic caused significant deviations in demand. Previous studies have also reported a steady increase in peak demand [61].

After data preprocessing and time series batching using the sliding window algorithm, 43 200 data points are available for training and testing. Table 1 describes the 65 input variables (J=65) used in this study and supplied to the model.

The multi-headed CNN is trained using data from the years 2015 to the end of 2018. Multivariate time series data from the last 7 days, collected at hourly intervals, is used to establish a 168-hour ($N = 168$) look-back window which serves as the input for multi-headed CNN. This data is then employed to generate an hourly forecast for the upcoming 24 h ($T = 24$). The multi-headed CNN is tested using 2019 data. Every test sample initiates at the day's onset, precisely at 00:00, covering a 24-hour period. This approach yields 356 non-overlapping test instances throughout the year. The multi-headed CNN parameters used in this study are listed in Tables 2–3.

The tests are performed using the following platforms; Python 3.10, Keras 2.8, Tensorflow 2.8.0, shap 0.41.0 and executed on an Intel i9-12900 CPU with 32 GB RAM.

Table 3
Training hyperparameters.

Parameter name	Parameter value
Batch size, Epochs, Steps per epoch	16, 300, 100
Learning rate	0.001 (Default)
Optimiser	Adam
Loss function	Mean Square Error (MSE)
Early stopping	Patience = 20, Monitor = val_loss
Validation split	0.1

Table 4
Elapsed computation time for XAI explanations.

Scope	Grad-CAM	SHAP
Local explanation	0.144 s	138 s
Global explanation	59 s	5580 s

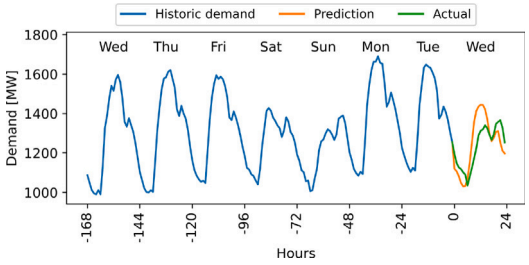


Fig. 3. National demand forecast for the holiday on Wednesday, 1 May 2019.

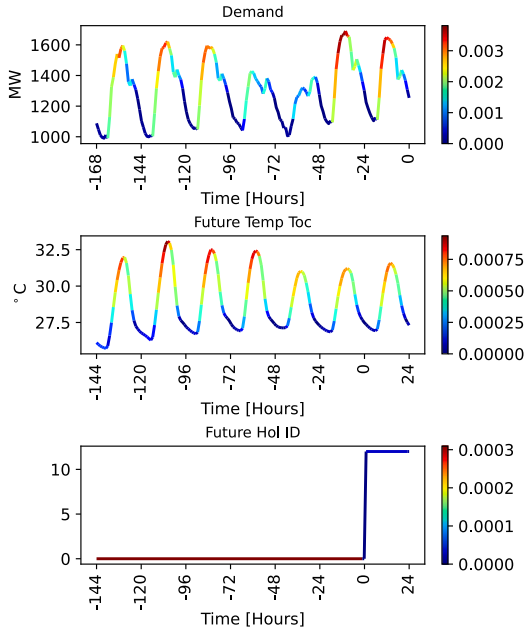
4.1. XAI results

Table 4 summarises the computation time needed for generating feature attributions using Grad-CAM and SHAP. Local and global explanations denote the computation time for generating feature attributions for single and multiple predictions, respectively, with the latter providing the feature rankings. The computational demand for global explanations is naturally higher due to the greater quantity of predictions. The Grad-CAM method proves to be computationally faster than the SHAP method for both local and global explanations, owing to its simpler computational complexity and reduced number of calculations. For instance, to generate an explanation for a single prediction, Grad-CAM only necessitates one forward propagation of input data through the model and a partial back-propagation of the gradients to the feature maps. Conversely, the SHAP method requires multiple iterations of model propagation using various input samples.

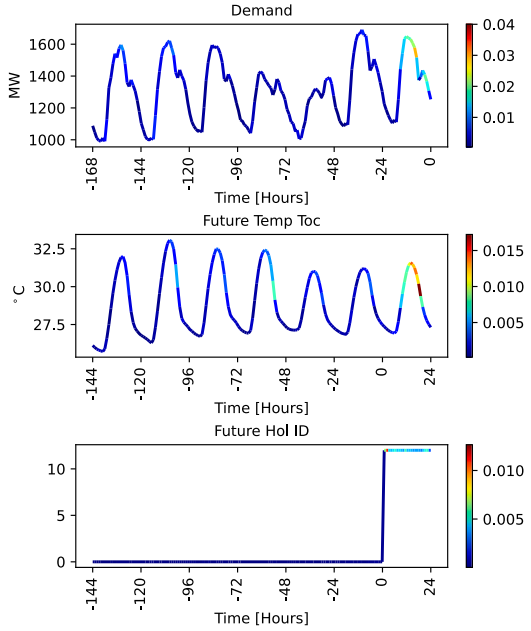
4.1.1. Local explanations

Fig. 3 showcases the predicted national demand for Wednesday, 1 May 2019, derived from a week's historical data encompassing 65 distinct features. On examining this figure, one can observe that both the actual and predicted demand profiles for this Wednesday closely mirror those usually seen on weekend, which are often indicative of reduced economic activities. Interestingly, this particular Wednesday was not just any weekday: it was National Labour Day in Panama, a public holiday. The noticeable dip in demand on this day indicates the influence of the public holiday. This observation aligns with the intuitive understanding that public holidays, much like weekends, lead to a slowdown in economic activities and, consequently, a reduction in demand.

The decision to focus on this specific day was deliberate. The contrast in demand, influenced by the public holiday, stands out clearly, making it an ideal candidate for analysis. Furthermore, this scenario provides a unique lens to explore the explanations offered by XAI. We anticipate that the XAI methods will emphasise the holiday feature, assigning it high attribution scores due to its evident impact. However,



(a) Grad-CAM



(b) SHAP

Fig. 4. Overlay of feature attribution on the input data, utilised for predicting national holidays.

while the holiday's influence is clear in this instance, it is essential to acknowledge that the dataset contains numerous other features. Although these other features may also significantly contribute to demand, their impact might not be as easily discernible solely from the profiles in Fig. 3

Figs. 4–5 provide a detailed view of the local feature attributions related to the prediction in Fig. 3, using Grad-CAM and SHAP. These feature attributions represent the local explanations for the predicted demand of the holiday. Fig. 4 is obtained by overlaying the feature attributions belonging to a specific feature onto the input data, which allows us to see which characteristics of the series were the most influential. The colour bar measures the importance of the feature at specific time steps. The time axis of "Future Temp Toc" and "Future Hol ID" is shifted to signify that theses feature contains future information. Fig. 5 abstracts the detail of the input data and provides an overview of feature attributions in the form of a heatmap. As all features' attribution scores are measured by the same colour bar, the heatmap representation makes it easy to compare the attributions of different features. Furthermore, these attributions are aligned with the original input data, which visualises the segments of the input data that influenced the model's prediction the most. It is important to mention that the heatmap for SHAP feature attributions displays the absolute SHAP values, which makes a visual comparison simpler. The bar chart in Fig. 6 displays the average feature attributions for the prediction of the holiday.

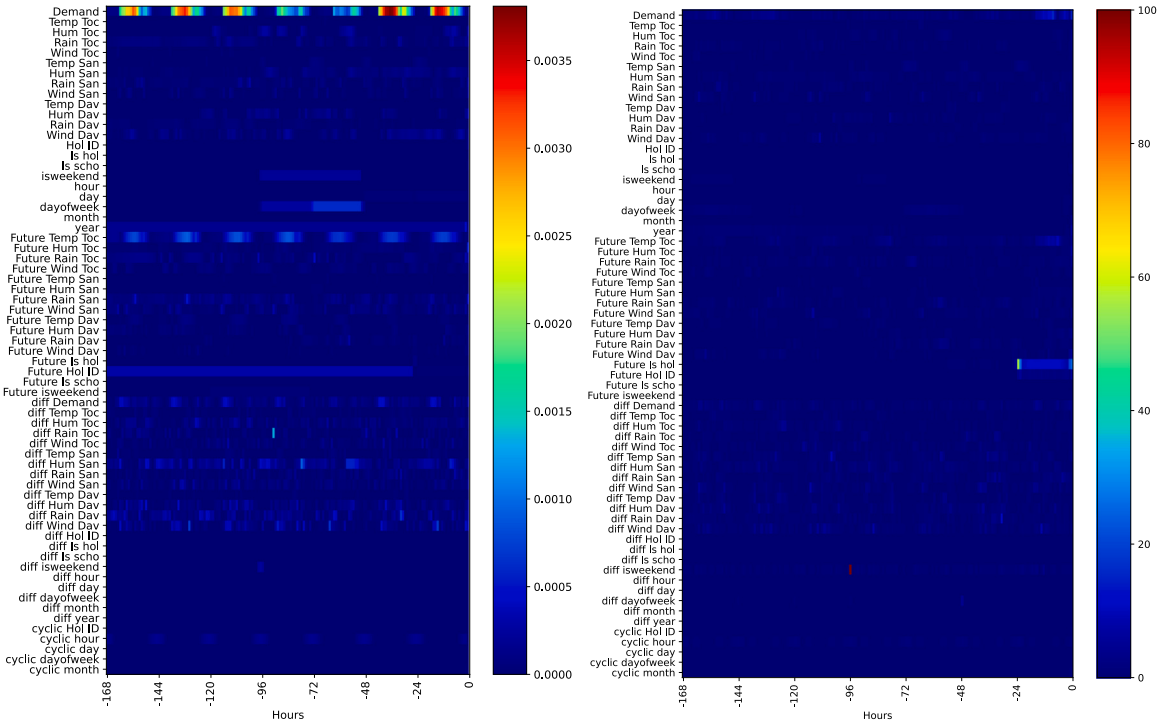
When the feature attribution scores produced by Grad-CAM are superimposed onto the original data in Fig. 4(a), it can be seen that high feature attributions for "Demand" are often associated with high values for demand. Notably, the highest attribution scores are assigned to the day with the highest demand. A similar trend is observed in "Future Temp Toc", where high temperatures coincide with high feature attributions. This trend extends to "isweekend", "dayofweek", "cyclic hour" and "diff isweekend".

Looking at Fig. 5, it can be seen that Grad-CAM highlights "Future Temp Toc" more than "Temp Toc". This signifies that the model considers future temperature to be more important than current and historic temperature records. This agrees with domain knowledge as the current temperatures are more likely to affect current demand. It thus follows that predicted temperatures affect predicted demand. However, a contradiction arises when looking at "Future Temp Toc" in Fig. 5(a), more closely. Despite the high attributions for "Future Temp Toc" shown in Fig. 5(a), in Fig. 4(a) temperatures from the past 4–6 days are assigned the highest attribution scores instead of future temperature, represented by hours 1–24. This inconsistency highlights a discrepancy in the reasoning provided by Grad-CAM which warrants further investigation. A closer look at the Grad-CAM explanations in Fig. 5(a) suggests that the holiday is not the sole feature influencing the prediction. However, without prior knowledge that the predicted demand corresponds to a holiday, pinpointing the exact reason for the dip in demand would arguably be a challenging endeavour.

In contrast, local explanations produced by SHAP, shown in Fig. 5(b), prominently displays the pronounced effect of national holiday on demand. SHAP also emphasises the significant influence of the most recent historical demand. Within these explanations, the temperatures forecasted for the next 1–24 h, denoted as "Future Temp Toc", are highlighted by SHAP as particularly impactful. Looking at the average attributions produced by SHAP in Fig. 6(b), "Future is hol" is the most significant feature. The insights derived from SHAP align closely with our intuitive understanding of the factors influencing demand.

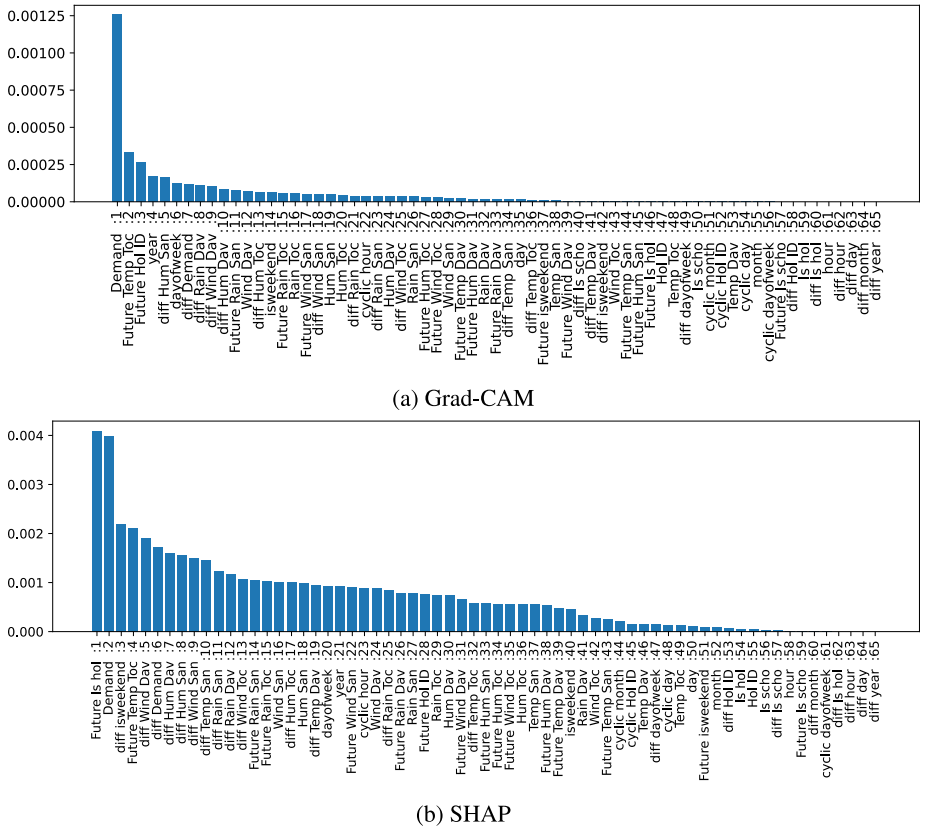
4.1.2. Global explanations

The feature rankings for both Grad-CAM and SHAP are illustrated in Fig. 7. These feature rankings display the global feature attributions, which are formed by aggregating individual feature attributions across all predictions made on the test set. This aggregation process captures the average importance of each feature and is displayed in Fig. 7 in descending order of importance. An inspection of the top 15 features ranked by both XAI methods, reveals a set of common features: "Demand", "diff Demand", "Future Temp Toc", "Future Temp Rain", "diff Rain Dav", and "year". It is observed that Grad-CAM assigns higher attribution scores to features with future information, while SHAP emphasises the differenced features. When it comes to identifying the most relevant features, the top 5–10 features as ranked by Grad-CAM in Fig. 7(a) seem to be the most significant, given that they constituent the largest proportion of the total attributions scores. However, when we evaluate the feature rankings of SHAP shown in Fig. 7(b), determining the threshold between relevant and irrelevant features becomes notably more challenging. Researchers in [58] separated features into groups



(a) Grad-CAM (b) SHAP

Fig. 5. Feature attribution heatmaps for the forecast of national holiday forecast on 1 May 2019.

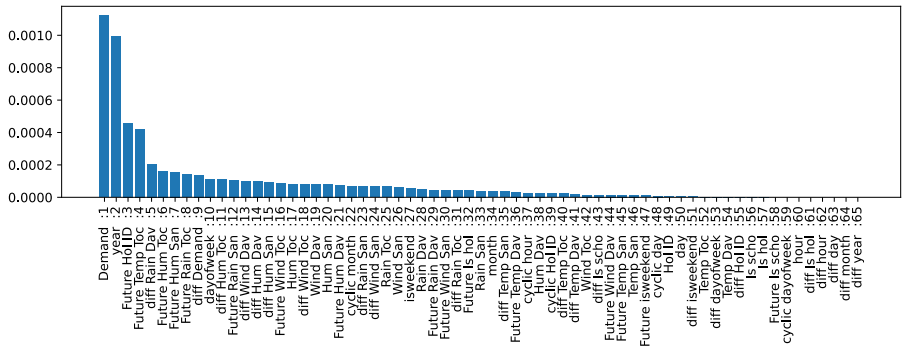


(a) Grad-CAM

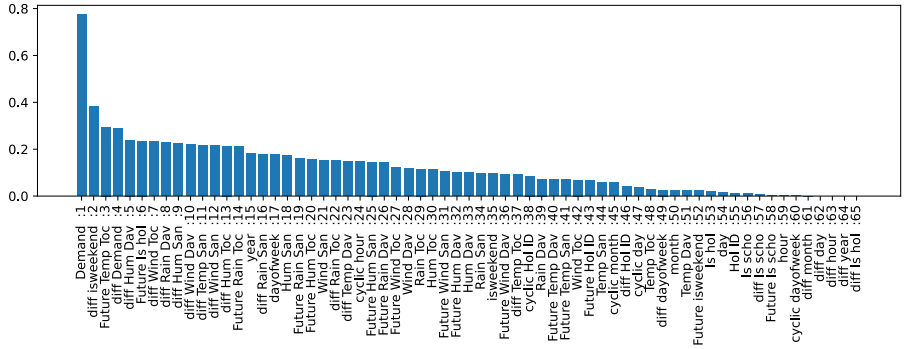
(b) SHAP

Fig. 6. Local average feature attributions scores that rank features in descending order of importance.

according to their attributions scores, however it is not clear how threshold that divides these groups were selected. This highlights a current limitation of XAI for feature selection, as there is a lack of a clear and widely accepted rule for setting this threshold.



(a) Grad-CAM



(b) SHAP

Fig. 7. Global average feature attributions scores that rank features in descending order of importance.

The ablation study in Fig. 8 indicates that discarding half of the lowest ranked features from both Grad-CAM and SHAP enhances forecasting accuracy and peak model performance is obtained when using the top 10 features from both XAI methods. This reinforces the idea that XAI methods are successful in pinpointing irrelevant features and consequently results in a tenfold reduction in model training time. The results also highlight how the inclusion of all features can degrade model performance.

In Fig. 8, a significant decrease in accuracy is observed when reducing the number of features from Grad-CAM's top 3 to top 2 features, suggesting that "Future Holiday ID" contributed significantly to forecasting accuracy. This implies that providing the model with information about upcoming holidays notably enhances the accuracy of forecasts. In comparison, the most significant drop in performance using SHAP feature rankings is observed when reducing the number of features from 7 to 5. This drop in model accuracy is likely due to the removal of "Future Is hol" given the demonstrated importance of the holiday feature. It is interesting to note that Grad-CAM assigned greater importance to the integer representation of the holiday feature whereas SHAP highlights the binary representation.

The multi-headed CNN trained using the top 10 to 33 features, as ranked by SHAP, generally outperforms those trained with features identified by Grad-CAM. This is evident from the lower average RMSE and MAPE shown in Fig. 8, suggesting superior forecasting accuracy. While the model seemingly achieved superior peak accuracy with the top 10 features from Grad-CAM, models trained using SHAP-selected features displayed greater consistency, evident from the reduced variance in their accuracy. This suggests that the model trained on features selected by SHAP results in better model stability and are less prone to overfitting than those trained on features identified by Grad-CAM. It is thus found that SHAP is better at identifying irrelevant features than Grad-CAM.

Despite the clear impact of holiday features on model accuracy, SHAP surprisingly ranks it as the 6th most important which ultimately leads to a premature ablation of the feature. This indicates that SHAP does not strictly rank features according to their predictive content. The same can be said about Grad-CAM where the "year" feature, ranked as the 2nd most significant feature, degrades model stability and contributes little in terms of predictive content. In Fig. 8 when 7 features remain, we see that SHAP feature rankings provide superior model accuracy compared to Grad-CAM, indicating that significant features are better preserved by SHAP during the feature elimination process. However this trend does not continue, as the top 3 and 5 features from Grad-CAM are shown to produce better performance. These findings underscore that neither XAI method provides a ranking of features strictly in order of their contribution to forecasting accuracy.

5. Discussion

A distinct advantage of XAI methods over wrapper methods is their computational efficiency, attributed to the elimination of the need for iterative model training required by wrapper methods. Grad-CAM demonstrates its computational efficiency in this high-dimensional application by providing feature rankings in 59 s, which based on previous studies [18], yields performance similar to that of filter methods. While our dataset only consists of 65 features, each feature carries 168 individual attributions. This amounts to a substantial total of 10 920 attributions. SHAP's computation for feature ranking takes roughly 100 times longer than Grad-CAM, requiring 1 h and 33 min to produce 10 920 attributions, which can be attributed to iterative perturbations of the model. To put this result in context, a study comparing wrapper methods found the best performing search algorithm identified the optimal feature set in 1778 s after 10 000 evaluations [22]. Given that the dataset in [22] comprises 1080 data points, in contrast to our dataset of 43 000 entries, it is reasonable to anticipate that a wrapper-based approach would necessitate a considerably longer duration in our case. SHAP's computational demands can be mitigated in various ways. Reducing the number of baseline and global attribution samples

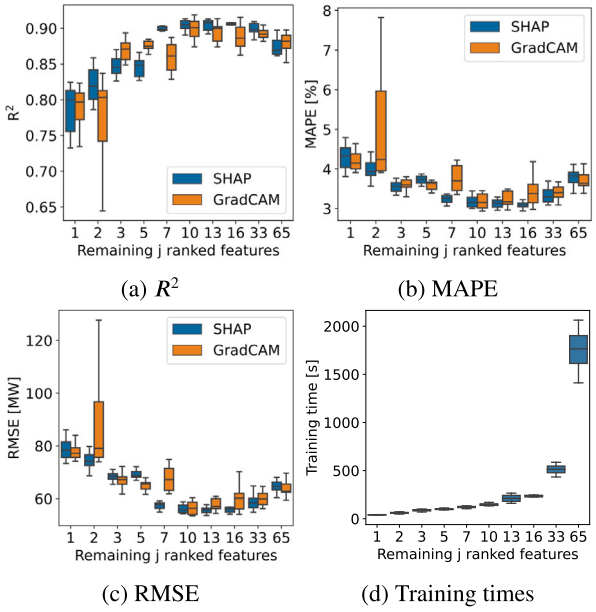


Fig. 8. Ablation study — Model response to retraining with reduced features.

can save time, provided the data distribution remains adequately represented to avoid bias. Additionally, shortening input lengths from 168 to 24 h, optimising the CNN architecture or selecting a simpler model can significantly improve SHAP's computational efficiency.

Computational efficiency alone would not be of much value if the XAI methods perform poorly as feature selection tools. Our experiments reveal that eliminating half of the irrelevant features identified by Grad-CAM and SHAP, lead to an improvement in forecasting accuracy and roughly 50% reduction in training time. For even better results, one can continue to prune lower-ranked features until there is a noticeable decline in the model's forecasting accuracy. While Grad-CAM's feature rankings yield model performance comparable to those chosen by SHAP, especially when considering the top 10 to 16 features, SHAP typically delivers superior model accuracy on average.

Using XAI for feature selection presents challenges similar to those found in filter and embedded methods that rank features by importance. Determining the boundary between relevant and irrelevant features becomes challenging when examining the feature ranking in 7. Without a universally accepted guideline for choosing the optimal number of features, one is compelled to continuously assess feature rankings to reach peak performance. While this method is computationally simple for our dataset, its practicality may be called into question for datasets containing a significantly larger number of features.

It is also interesting to analyse the variability in model performance during the ablation study. Since both Grad-CAM and SHAP are applied to the same model architecture and utilise the same training and testing data, any underlying differences can be attributed to the features selected by each method. Our findings indicate that features chosen by SHAP lead to more consistent model performance, as evidenced by the reduced variance in accuracy. This suggests that SHAP is better at identifying irrelevant features and may reduce the possibility of overfitting.

An important observation arises when abating higher ranked features. While removing the "Future Is hol" feature led to the most notable decline in model accuracy, SHAP ranked this feature as only the 6th most influential feature. It is not easy to pinpoint the precise cause for SHAP's inability to rank a crucial feature highly, however, we speculate that the selection of samples used to rank global importance might be the contributing factor. Given that the global feature attributions are an average magnitude of the local feature attributions, it is critical to select samples such that there is even an distribution of features. Fig. 5(b) aims to substantiate this point. There are 22 holidays in Panama out of the 365 days in a year. When using SHAP to interpret the forecast differences between a holiday and a regular weekday, it is evident that holiday features have no influence on a standard weekday forecast, yet they significantly impact the demand forecast on a holiday. If the holiday features are overshadowed by other features or outweighed by non-holiday samples, then the average becomes disproportionately skewed, leading to potentially misleading interpretations. This is a point of criticism of SHAP highlighted in a previous study [52]. Though the task of selecting an even distribution of features may be simple for discrete features, it is not clear how one should proceed for continuous features such as "temperature".

Grad-CAM also encounters issues with feature ranking. Contrary to assigning a low rank to a crucial feature, it ranked the "year" as the second most significant feature, even though it had minimal impact on forecasting accuracy. This is because, although the "year" feature is shown to contribute to the prediction in Fig. 5(a), its consistent average contribution might overshadow other features that have intermittent, yet critical, significance for specific forecasts. Intriguingly, the significance attributed to "year" does reflect the rising national demand of Panama. Yet, as demonstrated in Figs. 6(b)–7(b), SHAP suggests that variations in this feature exert only minimal effects on demand, in comparison to other factors.

It is important to note that Grad-CAM is sensitive to the scaling of data. Our initial tests indicate that when unscaled data is used, it appears to assign the highest attributions to features with the largest numerical values. Consequently, it not effective as a feature selection tool.

Another source of error in the feature rankings produced by Grad-CAM might stem from the fact that this method only measures positive influences on the forecast, neglecting factors that drive the magnitude of the forecast down. This is evident in Fig. 5(a) where low feature attributions are assigned to "Future Hol ID" or "Future is hol" between hours −24 and 0. This period of the features provides information to the model that the upcoming day will be a holiday. Such an oversight could result in a biased understanding of explanations and feature attributions. In [33], it was demonstrated that neglecting negative gradients can result in the inability to accurately identify certain faults in HVAC systems.

While XAI methods like SHAP and Grad-CAM assign feature attributions based on relevance to model predictions, these attributions are not direct indicators of predictive accuracy. This is observed in our ablation study, which shows that feature rankings do not strictly align with contributions to the model forecast accuracy. In this study, there is no assurance that the most effective feature combinations are necessarily among those ranked highest by either SHAP or Grad-CAM. Further research is warranted to truly assess the efficacy and optimal combinations of features.

Delving into local explanations offers insights not just about the rankings, but also about how individual features influence specific forecasts. When interpreting SHAP values, it is important to remember that they quantify the marginal contribution of that feature that led to the different between the prediction and the baseline. In the context of predicting the holiday on 1 May 2019, the feature "Future is hol" has the highest SHAP value, signifying its pivotal role in influencing the prediction relative to the baseline. Here, the baseline embodies the notion of an "average day". However, understanding this "average day" is nuanced. In our study, it is derived from the mean of demand profiles across the days sampled from the whole dataset. Given the composition of a week, where weekdays dominate with five out of the seven days, this baseline tends to resemble a standard weekday more than a weekend. This context underscores why the significant attribution to "Future is hol" corresponds to the observed reduction in demand on the holiday.

In contrast to SHAP, which quantifies a feature's impact relative to a specific baseline, we find that Grad-CAM offers a different perspective to the model's decision-making process. In Grad-CAM, feature attributions are computed through a weighted sum of the feature maps generated by the CNNs. These attributions are presumably anchored to a zero baseline, as they are not calculated relative to any baseline. To contextualise difference in perspectives provided by the XAI, we make use of the framework proposed by [64], which categorises various types of explanations. We find that Grad-CAM offers insights into the procedural mechanics of the model's prediction, akin to how multivariate linear regression explains the mathematical relationship between variables and the outcome. In short, Grad-CAM loosely explains how the model arrived at the predict and falls short in delivering causal explanations. For example, from Fig. 5(a), it is indeterminable what the cause of observed decline in demand is however it is evident that many feature played a role. Conversely, SHAP focuses on causal explanations rather than procedural explanations. It specifically highlights the influence of the holiday feature, aiding in understanding why a certain prediction was made.

6. Conclusion

This study examines Grad-CAM and SHAP as feature selection tools within the realm of multivariate time series load forecasting. Using XAI, we distil the multi-headed CNN's learned patterns into feature attributions, gauging their significance to predictions and facilitating feature selection by discarding low-attribution features. Furthermore, we have showcased that with the right array manipulations within the model pipeline, SHAP can be applied to explain even arbitrary model architectures.

Both Grad-CAM and SHAP have demonstrated their mettle by effectively identifying and eliminating irrelevant features from the model. This has led to tangible benefits such as improved forecast accuracy, reduced training time, and a diminished risk of overfitting. While Grad-CAM demonstrated its computational efficiency and its capability to identify features that frequently align with the performance benchmarks set by SHAP, several issues come to light. Not only does it manifest challenges in selecting features that produces consistent model performance, but it also does not offer intuitive explanations to the prediction of the model. Future studies can improve the Grad-CAM by accounting for factors that decrease predicted demand or explore the effects on the explanations by varying the kernel size of the CNN.

On the other hand, SHAP, despite its more intensive computational demands, consistently demonstrates an enhanced ability to identify pivotal features. Its strength lies in its capacity to offer intuitive and insightful explanations of the influencing factors behind forecasts, the selection of features that improve model performance, and the ability to be applied to any machine learning model. Nevertheless, there are times when even SHAP may not rank influential features with the precision one would expect. Future studies can attempt to improve SHAP's feature rankings by using sampling techniques that ensure an even distribution of features, especially when dealing with sparse features like holidays. Given the critical role that the baseline plays in interpreting SHAP values, there is an imperative need for establishing clear guidelines on the selection of an appropriate baseline set.

The explanations provided by the XAI are invaluable to the process of feature selection. Unlike traditional feature selection methods, Grad-CAM and SHAP provide context to the factors driving individual predictions. Since feature rankings are derived from aggregating attributions across all test set predictions, users gain a deeper understanding of the rationale behind these rankings. This transparency not only instills confidence but also enables users to apply expert judgement when necessary. In high-stakes domains, this level of insight is particularly vital, as relying solely on a model's overall test set performance can be misleading for assessing how it will behave in practice.

The potential of Grad-CAM to efficiently generate feature attributions, especially in applications with an exceptionally large set of features, holds great value and promises exciting developments in the field. One promising application is the forecasting of day-ahead electricity prices in European markets. This involves a wide array of features such as demand, energy production, electricity prices, weather patterns, forecasts and consumer behaviour across various countries [44]. The ability of Grad-CAM to handle such a vast range of features hints at its considerable potential for facilitating improvements in this area.

Looking ahead, further research is needed to refine these XAI methods and overcome the limitations identified in our study. In particular, there is a need for more effective methods for selecting the top number of features, and for further exploration of other XAI methods in time series forecasting domain. Future research could extend this work by contrasting the outcomes with wrapper-based methods to gauge the optimality of the features identified by XAI methods. Such a comparison would provide deeper insights into the efficacy of different feature selection strategies and offer a more nuanced understanding of the role and potential of XAI methods in model improvement and interpretability. Through this ongoing research, we can continue to unlock the full potential of these tools to foster more transparent, accurate, and interpretable forecasting models, paving the way for a brighter, more informed future.

CRediT authorship contribution statement

Corne van Zyl: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xianming Ye:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Raj Naidoo:** Resources, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xianming Ye reports financial support was provided by National Natural Science Foundation of China. Xianming Ye reports financial support was provided by National Research Foundation. Xianming Ye reports financial support was provided by Royal Academy of Engineering.

Data availability

Datasets related to this article can be found at [63], an open-source online data repository hosted at Mendeley Data.

Acknowledgements

This research is jointly supported by National Key R&D Program of China (Grant No. 2021YFE0199000), National Natural Science Foundation of China (Grant No. 62133015), National Research Foundation China/South Africa Research Cooperation Programme, China/South Africa Bilateral with Grant No. 148762, and Royal Academy of Engineering Transforming Systems through Partnership grant scheme, UK with reference No. TSP2021\100016.

References

[1] Kazmi H, Tao Z. How good are TSO load and renewable generation forecasts: Learning curves, challenges, and the road ahead. Appl Energy 2022;323:119565. http://dx.doi.org/10.1016/j.apenergy.2022.119565.

[2] Bouktif S, Fiaz A, Ouni A, Serhani MA. Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. Energies 11(7). http://dx.doi.org/10.3390/en11071636.

[3] Aslam S, Herodotou H, Mohsin SM, Javaid N, Ashraf N, Aslam S. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. Renew Sustain Energy Rev 2021;144:110992. http://dx.doi.org/10.1016/j.rser.2021.110992.

[4] Hong T, Pinson P, Wang Y, Weron R, Yang D, Zareipour H. Energy forecasting: A review and outlook. IEEE Open Access J Power Energy 2020;7:376–88. http://dx.doi.org/10.1109/OAJPE.2020.3029979.

[5] Machlev R, Heistrene L, Perl M, Levy K, Belikov J, Mannor S, et al. Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy AI 2022;9:100169. http://dx.doi.org/10.1016/j.egyai.2022.100169.

[6] Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018;73:1–15. http://dx.doi.org/10.1016/j.dsp.2017.10.011.

[7] Covert I, Lundberg SM, Lee S-I. Understanding global feature contributions with additive importance measures. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. Advances in neural information processing systems, vol. 33. Curran Associates, Inc.; 2020, p. 17212–23.

[8] Almuallim H, Dietterich TG. Learning with many irrelevant features. In: Proceedings of the ninth national conference on artificial intelligence - volume 2. AAAI Press; 1991, p. 547–52, URL https://api.semanticscholar.org/CorpusID:12494914.

[9] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 2004;5:1205–24.

[10] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40(1):16–28. http://dx.doi.org/10.1016/j.compeleceng.2013.11.024, 40th-year commemorative issue.

[11] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.

[12] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17. http://dx.doi.org/10.1093/bioinformatics/btm344.

[13] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 2019;267:1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.

[14] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70–9. http://dx.doi.org/10.1016/j.neucom.2017.11.077.

[15] Eseye AT, Lehtonen M, Tukia T, Uimonen S, John Millar R. Machine learning based integrated feature selection approach for improved electricity demand forecasting in decentralized energy systems. IEEE Access 2019;7:91463–75. http://dx.doi.org/10.1109/ACCESS.2019.2924685.

[16] Effrosynidis D, Arampatzis A. An evaluation of feature selection methods for environmental data. Ecol Inform 2021;61:101224. http://dx.doi.org/10.1016/j.ecoinf.2021.101224.

[17] Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends 2020;1(2):56–70. http://dx.doi.org/10.38094/jastt1224, URL https://jastt.org/index.php/jasttpath/article/view/24.

[18] Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Statist Data Anal 2020;143:106839. http://dx.doi.org/10.1016/j.csda.2019.106839.

[19] Li W, Becker DM. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling. Energy 2021;237:121543. http://dx.doi.org/10.1016/j.energy.2021.121543.

[20] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97(1):273–324. http://dx.doi.org/10.1016/S0004-3702(97)00043-X.

[21] Xue B, Zhang M, Browne WN, Yao X. A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput 2016;20(4):606–26. http://dx.doi.org/10.1109/TEVC.2015.2504420.

[22] Agrawal P, Abutarboush HF, Ganesh T, Mohamed AW. Metaheuristic algorithms on feature selection: A survey of one decade of research (2009–2019). IEEE Access 2021;9:26766–91. http://dx.doi.org/10.1109/ACCESS.2021.3056407.

[23] Nguyen BH, Xue B, Zhang M. A survey on swarm intelligence approaches to feature selection in data mining. Swarm Evol Comput 2020;54:100663. http://dx.doi.org/10.1016/j.swevo.2020.100663.

[24] Zheng H, Yuan J, Chen L. Short-term load forecasting using EMD-LSTM neural networks with a XGBoost algorithm for feature importance evaluation. Energies 10(8). http://dx.doi.org/10.3390/en10081168.

[25] Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinform 2008;9(1):307. http://dx.doi.org/10.1186/1471-2105-9-307.

[26] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 2018;58(1):267–88. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[27] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 2003;53(1):23–69. http://dx.doi.org/10.1023/A:1025667309714.

[28] Niu T, Wang J, Lu H, Yang W, Du P. Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting. Expert Syst Appl 2020;148:113237. http://dx.doi.org/10.1016/j.eswa.2020.113237.

[29] Guyon IM, Weston J, Barnhill SD, Vapnik VN. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.

[30] Theissler A, Spinnato F, Schlegel U, Guidotti R. Explainable AI for time series classification: A review, taxonomy and research directions. IEEE Access 2022;10:100700–24. http://dx.doi.org/10.1109/ACCESS.2022.3207765.

[31] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision. 2017, p. 618–26. http://dx.doi.org/10.1109/ICCV.2017.74.

[32] Kim T-Y, Cho S-B. Predicting residential energy consumption using CNN-LSTM neural networks. Energy 2019;182:72–81. http://dx.doi.org/10.1016/j.energy.2019.05.230.

[33] Li G, Yao Q, Fan C, Zhou C, Wu G, Zhou Z, Fang X. An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems. Build Environ 2021;203:108057. http://dx.doi.org/10.1016/j.buildenv.2021.108057.

[34] Machlev R, Perl M, Belikov J, Levy KY, Levron Y. Measuring explainability and trustworthiness of power quality disturbances classifiers using XAI—explainable artificial intelligence. IEEE Trans Ind Inf 2022;18(8):5127–37. http://dx.doi.org/10.1109/TII.2021.3126111.

[35] Assaf R, Giurgiu I, Bagehorn F, Schumann A. MTEX-CNN: Multivariate time series explanations for predictions with convolutional neural networks. In: 2019 IEEE international conference on data mining. 2019, p. 952–7. http://dx.doi.org/10.1109/ICDM.2019.00106.

[36] Li Y, Yang H, Li J, Chen D, Du M. Eeg-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM. Neurocomputing 2020;415:225–33. http://dx.doi.org/10.1016/j.neucom.2020.07.072.

[37] Fauvel K, Lin T, Masson V, Fromont E, Termier A. XCM: An explainable convolutional neural network for multivariate time series classification. Mathematics 9(23). http://dx.doi.org/10.3390/math9233137.

[38] Ribeiro MT, Singh S, Guestrin C. Why should i trust you? explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 1135–44.

[39] Veerappa M, Anneken M, Burkart N, Huber MF. Validation of XAI explanations for multivariate time series classification in the maritime domain, journal of computational. Science 2022;58:101539. http://dx.doi.org/10.1016/j.jocs.2021.101539.

[40] Schlegel U, Arnout H, El-Assady M, Oelke D, Keim DA. Towards a rigorous evaluation of XAI methods on time series. In: 2019 IEEE/CVF international conference on computer vision workshop. Los Alamitos, CA, USA: IEEE Computer Society; 2019, p. 4197–201. http://dx.doi.org/10.1109/ICCVW.2019.00516.

[41] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems, vol. 30. Curran Associates, Inc.; 2017, p. 4765–74.

[42] Chakraborty D, Alam A, Chaudhuri S, Başağaoğlu H, Sulbaran T, Langar S. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. Appl Energy 2021;291:116807. http://dx.doi.org/10.1016/j.apenergy.2021.116807.

[43] Mitrentsis G, Lens H. An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. Appl Energy 2022;309:118473. http://dx.doi.org/10.1016/j.apenergy.2021.118473.

[44] Tschora L, Pierre E, Plantevit M, Robardet C. Electricity price forecasting on the day-ahead market using machine learning. Appl Energy 2022;313:118752. http://dx.doi.org/10.1016/j.apenergy.2022.118752.

[45] Białek J, Bujalski W, Wojdan K, Guzek M, Kurek T. Dataset level explanation of heat demand forecasting ANN with SHAP. Energy 2022;261:125075. http://dx.doi.org/10.1016/j.energy.2022.125075.

[46] Ozyegen O, Ilic I, Cevik M. Evaluation of interpretability methods for multivariate time series forecasting. Appl Intell 2022;52:4727–43. http://dx.doi.org/10.1007/s10489-021-02662-2.

[47] Lundberg S, et al. SHAP (shapley additive explanations). 2023, https://shap.readthedocs.io/en/latest/index.html. [Accessed 18 May 2023].

[48] Haug J, Zurn S, El-Jiz P, Kasneci G. On baselines for local feature attributions. 2021, arXiv abs/2101.00905.

[49] Sturmfels P, Lundberg S, Lee S-I. Visualizing the impact of feature attribution baselines. Distill 2020;5(1). http://dx.doi.org/10.23915/distill.00022.

[50] Chen H, Lundberg SM, Lee S-I. Explaining a series of models by propagating shapley values. Nature Commun 2022;13(1):4512. http://dx.doi.org/10.1038/s41467-022-31384-3.

[51] Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell 2021;298:103502. http://dx.doi.org/10.1016/j.artint.2021.103502.

[52] Fryer D, Strümke I, Nguyen H. Shapley values for feature selection: The good, the bad, and the axioms. IEEE Access 2021;9:144352–60. http://dx.doi.org/10.1109/ACCESS.2021.3119110.

[53] Marcílio WE, Eler DM. From explanations to feature selection: assessing shap values as feature selection mechanism. In: 2020 33rd SIBGRAPI conference on graphics, patterns and images. 2020, p. 340–7. http://dx.doi.org/10.1109/SIBGRAPI51738.2020.00053.

[54] Ghadimi N, Akbarimajd A, Shayeghi H, Abedinia O. Two stage forecast engine with feature selection technique and improved meta-heuristic algorithm for electricity load forecasting. Energy 2018;161:130–42. http://dx.doi.org/10.1016/j.energy.2018.07.088.

[55] Dai Y, Zhao P. A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization. Appl Energy 2020;279:115332. http://dx.doi.org/10.1016/j.apenergy.2020.115332.

[56] González-Vidal A, Jiménez F, Gómez-Skarmeta AF. A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. Energy Build 2019;196:71–82. http://dx.doi.org/10.1016/j.enbuild.2019.05.021.

[57] Wang Y, Sun S, Chen X, Zeng X, Kong Y, Chen J, et al. Short-term load forecasting of industrial customers based on svmd and xgboost. Int J Electr Power Energy Syst 2021;129:106830. http://dx.doi.org/10.1016/j.ijepes.2021.106830.

[58] Sim T, Choi S, Kim Y, Youn SH, Jang D-J, Lee S, et al. Explainable ai (xai)-based input variable selection methodology for forecasting energy consumption. Electronics 2022;11(18). http://dx.doi.org/10.3390/electronics11182947.

[59] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2012;60:84–90, URL https://api.semanticscholar.org/CorpusID:195908774.

[60] Yazici I, Beyca OF, Delen D. Deep-learning-based short-term electricity load forecasting: A real case application. Eng Appl Artif Intell 2022;109:104645. http://dx.doi.org/10.1016/j.engappai.2021.104645.

[61] Ibrahim B, Rabelo L. A deep learning approach for peak load forecasting: A case study on Panama. Energies 14(11). http://dx.doi.org/10.3390/en14113039.

[62] Lim B, Arık SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast 2021;37(4):1748–64. http://dx.doi.org/10.1016/j.ijforecast.2021.03.012.

[63] Aguilar Madrid E. Short-term electricity load forecasting (Panama case study). 2021, http://dx.doi.org/10.17632/byx7sztj59.1, version: V1.

[64] Cabitza F, Campagner A, Malgieri G, Natali C, Schneeberger D, Stoeger K, et al. Quod erat demonstrandum? - towards a typology of the concept of explanation for the design of explainable ai. Expert Syst Appl 2023;213:118888. http://dx.doi.org/10.1016/j.eswa.2022.118888.